Concept Maps: Theory, Methodology, Technology
Proc. of the First Int. Conference on Concept Mapping
A. J. Cañas, J. D. Novak, F. M. González, Eds.
Pamplona, Spain 2004

C-TOOLS AUTOMATED GRADING FOR ONLINE CONCEPT MAPS WORKS WELL WITH A LITTLE HELP FROM "WORDNET"

Scott H. Harrison[1], Joshua L. Wallace[2], Diane Ebert-May[3] And Douglas B. Luckie[2,4]
[1]Department of Microbiology and Molecular Genetics, [2]Lyman Briggs School of Science, [3]Department of Plant Biology, and [4]Department of Physiology, Michigan State University
Email: harris41@msu.edu, ctools.msu.edu

Abstract. Criterion concept maps developed by instructors in the C-TOOLS project contain numerous expert-generated or "correct" propositions manually created by expert users connecting two object or subject phrases together with a linking phrase. WordNet, an electronic lexical database, was then used to construct additional proposition derivatives by supplying different linking phrases in place of those originally specified by users. Derived propositions were made from original propositions by substituting linking phrase verbs with troponyms, synonyms, and antonyms. During the past year some 1298 students created concept maps (with 35404 propositions) aided by automatic grading and the WordNet® propositions. We have now studied how successful WordNet was at creating valid additional linking words like those generated by experts. By comparing manual assessments of derived propositions to manual assessments of original propositions, the persistence of correctness was evaluated for determinative factors such as frequencies and senses of usage. Results from data analyses are compared to parameters for WordNet usage in order to better determine the potential for the refining of concept map grading algorithms. An empirical approach to word sense relationships is presented as an iterating step to progressively prototype and test optimizations.

1 Introduction

Automatic grading features associated with the C-TOOLS project's Concept Connector began in 2003 to amplify instructor-designed grading matrices with synonyms from WordNet® (Fellbaum, Ed., 1998). At present, for 35404 propositions, 9211 can be evaluated by an automated grading mechanism called Robograder™. WordNet-powered amplification enabled 971 of the 9211 propositions to be evaluated when the existing grading matrices would not otherwise make an assessment. Currently, Robograder™ indiscriminately accepts linking phrase synonyms independent of frequency and word sense.

Visual examinations of automatically graded maps indicated few false positives or false negatives despite Robograder's treatment of multiple synsets as interchangeably equivalent. In theory amplifying grading rubrics by using the superset of all available synonyms should introduce errors into rubrics since multiple and conflicting meanings often exist. One explanation for the observed success of indiscriminate acceptance of synonyms is that users may more likely choose words within a relevant set of synonyms (known in WordNet as a "synset"). Thus, when developing concept maps, users appear to be less inclined to take "stabs in the dark" than compared to a conceptual strategy which favors semantically plausible word choices. An existing limitation to our WordNet integration concerns the grammatically parsing of linking phrases. The goal of parsing is to evaluate the meaning of the central verb of the linking phrase while accounting for adverbs, articles, adjectives, and prepositions. Currently, our approach is to evaluate linking phrases that just consist of a single word that WordNet recognizes as a verb.

We present a survey of how the assessment of a proposition's correctness changes when making related substitutions involving synonyms, troponyms, or antonyms. Different lexical substitutions were employed to help relate changes in linking phrases to factors of ambiguity and correctness, characteristics used in deductive models concerning WordNet and concept maps (Cañas, Valerio, Lalinde-Pulido, Carvalho, & Arguedas, 2003). Artificial creation of new propositions is an approach to preemptively identify constraints and ranges on linking word choices not yet made by users and are studied within the empirical context of a random sampling approach. The joining of theory with real-world feedback helps further deploy a design experiment framework to improve learning through the use of web-based concept maps (Luckie, Harrison, & Ebert-May, 2004).

2 Analysis of Original and Derived Propositions

2.1 Data Acquisition

There were 35404 propositions available from Michigan State University's C-TOOLS server. A random sample of 250 propositions was gathered and divided into 5 separate sets of 50 each. Each of these original propositions consisted of a starting concept phrase, a linking phrase, and a terminal concept phrase.

2.2 Manual Assessment

Manual assessment of propositions was done by hand without aid of electronic references or algorithms. Manual assessments of the 5 sets were performed by the first two authors. Scorings for each proposition were: 1 (correct, e.g. “Photosynthesis - needs - Carbon dioxide”), X (incorrect, e.g. “DNA - transcribes – RNA”), 0 (ambiguous, e.g. “atom - is made of – neutron”), and S (structural violation, e.g. “ocans - evaporation – atmosphere”). Structural violations were for propositions with grammar problems such as spelling errors and linking phrases that do not contain a verb. Ambiguous scores were given to propositions that could only be scored as correct when viewed in a reasonably plausible context of surrounding propositions.

2.3 WordNet Derivatives

Version 2.0 of the software database WordNet was used to generate "proposition derivatives" by making linking phrase substitutions based on WordNet's thesaurus-like lexical capabilities. There were two criteria for the generation of proposition derivatives. First, derived propositions were made from linking phrases consisting of a single verb. Only 121 of the 250 original propositions met this single verb word criterion. Second, at minimum, the WordNet database had to have three available choices per lexical relationship. An original proposition’s linking verb could thus have up to nine derivatives (i.e. 3 antonyms, 3 troponyms, and 3 synonyms). Each triplet, as generated per lexical relationship (e.g. three antonyms), is called a trio. Trios provide an initial comparative range of data concerning the sense of usage and polysemy counts specific to both lexically similar and dissimilar derivations made from original propositions. With the single verb and triplet criteria, WordNet enabled us to construct 30 antonym derivatives, 243 troponym derivatives, and 234 synonym derivatives per proposition. Grading of proposition derivatives was delegated by the originating proposition sets. Graders A and B both graded derivative set 5. Grader A graded derivative sets 3 and 4. Grader B graded derivative sets 1 and 2.

2.4 Data Analysis

The manual assessments of original and derivative propositions were scrutinized in order to both summarize and make insights into relationships that may concern automated strategies of assessment. Assessments of original and derived propositions were enumerated in order to show relative ratios of correctness, ambiguity, and grammatical errors. Trios were analyzed for fluctuations in correctness and incorrectness. Trends between correctness and polysemy were investigated.

Graders A and B had reproducible similarity to their scoring patterns as determined by the Kappa statistic (Cohen, 1960). The Kappa statistic ($\kappa = 0.552$) was calculated with $p_o = 0.720$ and $p_e = 0.374$ suggesting good reproducibility ($0.4 \leq \kappa \leq 0.75$). The level of significance for this degree of association is < 0.10. For the manual assessments of the 250 original propositions, 72% of the assessments between the two graders were identical (180 propositions). Opposite assessments of correctness (1 versus X) occurred 5.6% of the time. Remaining differences for the assessment of individual propositions were primarily attributable to issues unrelated to exacting qualifications of correctness. For example, 30 instances of disagreement involved only one grader assigning an S score and 26 instances of disagreement involved one grader cautiously assigning a 0 (ambiguous) score in contrast to 1 or X scorings. While our approach has statistically significant repeatability for scoring ratio properties and strong consistency for exacting qualifications of proposition correctness, further refinement would involve better synchronization between graders’ approaches to assumptions of context and handling of grammatical logistics. When looking at the jointly graded WordNet derivative set ($n = 144$), the agreement between grader A and grader B was 70% ($p_o = 0.701$). The degree of association is just marginally reproducible based on $\kappa = 0.375$ and this reduction may be attributable to fewer shared contextual assumptions between graders due to loss of the original word choice. Scoring dynamics appear to be conserved; joint scorings for derivatives rise in agreement when considering just 1 and X scores, and the κ value does not suggest complete insignificance ($\alpha = 0.13$).

Antonym derivatives were found to be always incorrect. Of 21 antonyms graded by grader A, 21 were graded as incorrect. Of the 18 antonyms graded by grader B, 18 were graded as incorrect. Assessments for original, synonym-derived, and troponym-derived propositions are shown in Table 1 and encompass a range of assessment across all four grading categories (1, 0, X, and S). When the range of assessment is limited to 1 and X, grader A found 25.6% of synonym-derived propositions to be correct and 16.8% of troponym-derived propositions to be correct. For 1 and X scorings, grader B found 43.5% of synonym-derived propositions to be correct and 31.7% of troponym-derived propositions to be correct.

Score	Original propositions		Synonym-derived propositions		Troponym-derived propositions	
	Grader A	Grader B	Grader A	Grader B	Grader A	Grader B
Correct	141	123	32	68	21	53
Incorrect	12	32	93	88	104	114
Ambiguous	33	13	16	0	22	0
Structural violation	64	82	0	0	0	1

Table 1: Summary of manual assessment scores for original, synonym-derived and troponym-derived propositions.

The construction of trios involves random sampling from each WordNet-generated set of antonyms, synonyms, and troponyms. If conflicting meanings inside each set cause a general variation of proposition correctness, then clustering of correct or incorrect assessments within trios should not differ from a distribution of correct assessments that is random with respect to triplet structure. For the 57 synonym-derived trios assessed by grader A and the 81 synonym-derived trios assessed by grader B, the distributions showed no significant difference ($\chi^2 = 1.59$, $p = 0.66$ and $\chi^2 = 2.07$, $p = 0.56$ respectively). For the 54 troponym-derived trios assessed by grader A and the 71 troponym-derived trios assessed by grader B, the distributions also showed no significant difference ($\chi^2 = 2.64$, $p = 0.45$ and $\chi^2 = 5.57$, $p = 0.13$ respectively).

The general variability of correctness occurring within trios was investigated further by measuring how assessment score changes relate to similarities in meaning for derived proposition linking verbs. The WordNet database organizes lexical sets into subsets (termed "synsets") grouped together by similar meaning. Pairs of propositions occurring within trios were analyzed for having dissimilar correctness scores 1 and X, and for whether each proposition's linking verb was a member of the same synset. Shared synset membership for troponym derivatives occurred for 67% (grader A) and 49% (grader B) of all trio pairings that had an assessment score transition from 1 to X. Scoring transitions from 1 to X were next contrasted to within-trio proposition pairs where both propositions were assessed with a score of 1. Shared synset membership for troponym derivative pairs occurred for 100% (grader A) and 81% (grader B) of all such trio pairings that had a common assessment score of 1. Synonym derivatives were analyzed in similar fashion. Shared synset membership for synonym derivatives occurred for 15% (grader A) and 14% (grader B) of all trio pairings that had an assessment transition from 1 to X. Shared synset membership for synonym derivative pairs occurred for 27% (grader A) and 31% (grader B) of all such trio pairings that had a common assessment score of 1. Thus, for both troponyms and synonyms, membership of two verbs in the same synset implicates retained assessments of correctness.

Correctness was then analyzed for its impact on polysemy count distributions. For original propositions, polysemy distribution values were $\mu = 3.43$, $\sigma = 7.72$ and $\mu = 4.08$, $\sigma = 6.46$ for incorrect and correct propositions respectively. For derived propositions, polysemy distribution values were $\mu = 6.01$, $\sigma = 7.51$ and $\mu = 8.90$, $\sigma = 11.40$ for incorrect and correct propositions respectively. The increase of polysemy counts with correctness was attributed both to moderately high polysemy count ranges (>20) corresponding to the correctness of propositions by a factor of 2.4 and to a distinct trend for low polysemy count ranges (<5) corresponding to a 10% rise in the incorrectness of propositions.

3 Summary

Accuracy of correctness within synonym and troponym-derived trios was distributed randomly implying general factors that influence variance in meaning within lexical sets. Similar meanings between linking verbs positively associated with conserving correctness. High polysemy counts had some correspondence with correctness implying that correct propositions are more likely to occur for common words. The implication of common words with correctness may further relate to studies that show learning to occur favorably within the context of either experienced usage of a term (Novak, 1990) or familiarity with a term (Wittrock, 1992).

As a design experiment, the focus of C-TOOLS is to form a solution that works with data in real classrooms (Collins, Joseph, & Bielaczyc, 2004). An approach that amplifies correctness across multiple synsets appeared to work on concept maps made by real users. Such an indiscriminating approach was faulty when applied to randomly generated sets of synonyms and troponyms. Thus, the data supports that synonyms and troponyms can be used as sets for further identifying both correct and incorrect propositions. Although it may appear that there is only a 10% gain by using synonyms for automatic grading, this is only from the standpoint of

automating the assessment at the proposition level. At the larger concept map level, there are highly interconnected concept words that follow a pattern of classroom consensus and also correspond to student performance (Luckie, Harrison, & Ebert-May, 2004). Better understanding of the linking words around major hubs would aid us to analyze the formative dynamics of how users in a classroom interconnect concepts and, potentially, knowledge domains. Analysis and further improvements to Robograder™ cannot just be limited to synset hierarchies of each individual linking word since there are content-dependent dynamics of semantic overlap that influence how words can sensibly connect to other words (Banerjee & Pedersen, 2003).

A tactical challenge for manually assessing propositions is to establish reproducible standards that can be further refined. For this study, the treatment was organized by a custom-designed statistical query language that would scale well for much larger lists of user propositions. To aid further refinement of analyses, our current query algorithms are being packaged into a module for use with the R project (Ihaka & Gentleman, 1996). As an emerging, openly usable tool, the C-TOOLS project uses free resources like WordNet and R to have a cross-institutional scope for the building of experiments to help design effective concept map assessment algorithms.

4 Acknowledgements

This research project is supported by grant DUE 0206924 from the National Science Foundation. We thank Drs. Janet Batzli, Susan Bagley, James Smith, Lynmarie Posey, Walter Benenson, Michele Ouelette, Duncan Sibley, Tammy Long and Deborah Linton for their assistance in the implementation of this project.

5 References

Banerjee, S., & Pedersen, T. (2003). *Extended Gloss Overlaps as a Measure of Semantic Relatedness*. Paper presented at IJCAI 2003 – 18th International Joint Conference on Artificial Intelligence.

Cañas, A. J., Valerio, A., Lalinde-Pulido, J., Carvalho, M., & Arguedas, M. (2003). *Using WordNet for Word Sense Disambiguation to Support Concept Map Construction*. Paper presented at SPIRE 2003 – 10th International Symposium on String Processing and Information Retrieval.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement,* 20:37–46.

Collins, A., Joseph, D. & Bielaczyc, K. (2004). Design research: Theoretical and methodological issues. *Journal of the Learning Sciences, 13*(1), 15–42.

Fellbaum, C. Ed. (1998). *WordNet – An Electronic Lexical Database,* MA: MIT Press.

Ihaka, R., & Gentleman R. (1996). R: A Language for Data Analysis and Graphics, *Journal of Computational and Graphical Statistics, 5*, 299-314.

Luckie, D. B., Harrison, S. H., & Ebert-May, D. (2004). *Introduction to C-TOOLS: Concept Mapping Tools for Online Learning*. Paper presented at CMC 2004 – 1st International Conference on Concept Mapping.

Novak, J. (1990). Concept Mapping: A Useful Tool for Science Education. *Journal of Research in Science Teaching, 27*(10), 937-949.

Wittrock, M. C. (1992). Generative Learning Processes of the Brain. *Educational Psychologist, 27*(4), 531-541.